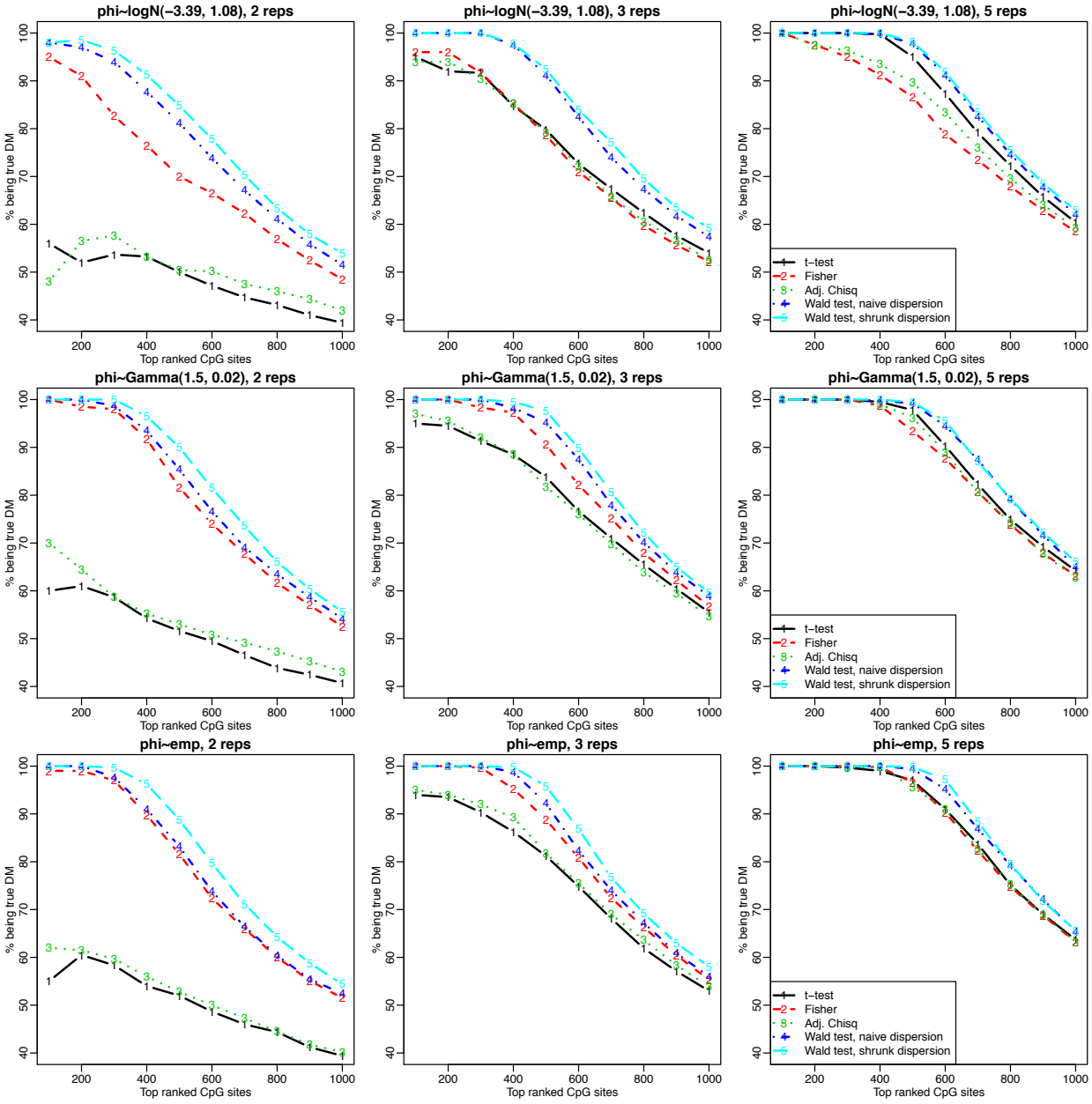
A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data

Hao Feng[1,2], Karen N. Conneely[2*], Hao Wu[1*]

Supplemental Material 1

1. Comparison of DML detection accuracies —part A

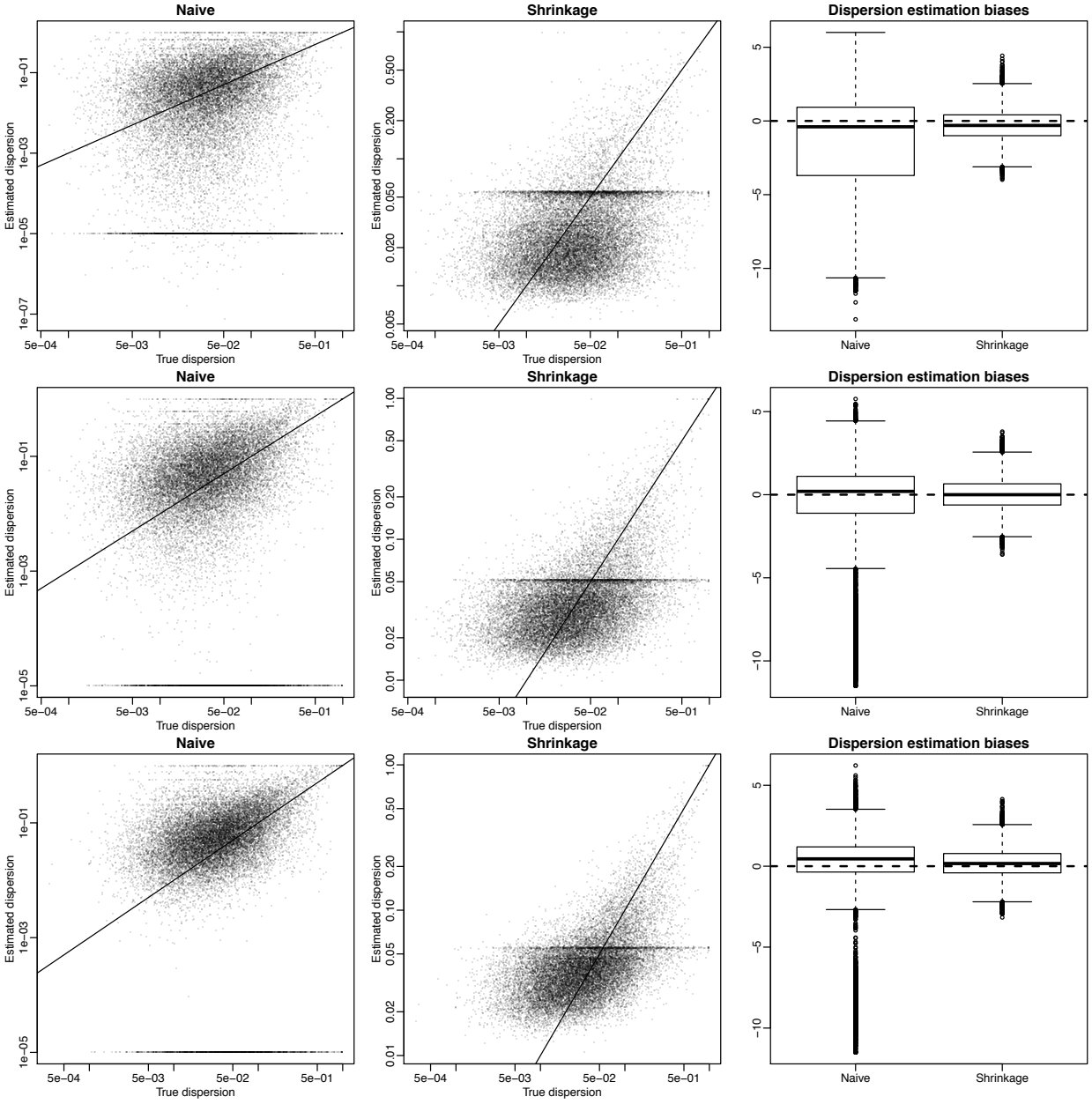
We provide results for additional simulations comparing the accuracy of DML detection using different methods. Supplementary Figure 1 provides simulation results based on the mouse genome (Smith *et al.*, 2012, Gene Expression Omnibus accession GSE34864), where the number of replicates and dispersion distributions used in the simulations vary.



Supplementary Figure 1. The True Discovery Rate (TDR) plot of DML detection accuracy from our proposed method for the top 1000 CpG sites. After testing for DML, CpG sites are ranked by p-value from low to high. The proportion of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. The dispersion parameter Ø comes from the log-normal distribution (1st row), Gamma distribution (2nd row), and empirical distribution from the real data (3rd row). In situations where replicate numbers are low (1st column), our proposed method shows the largest improvement since it accounts for the difficulty of estimating variance from 2 replicates per group. Our proposed Wald test with shrunk dispersion estimator performs the best across all nine simulations.

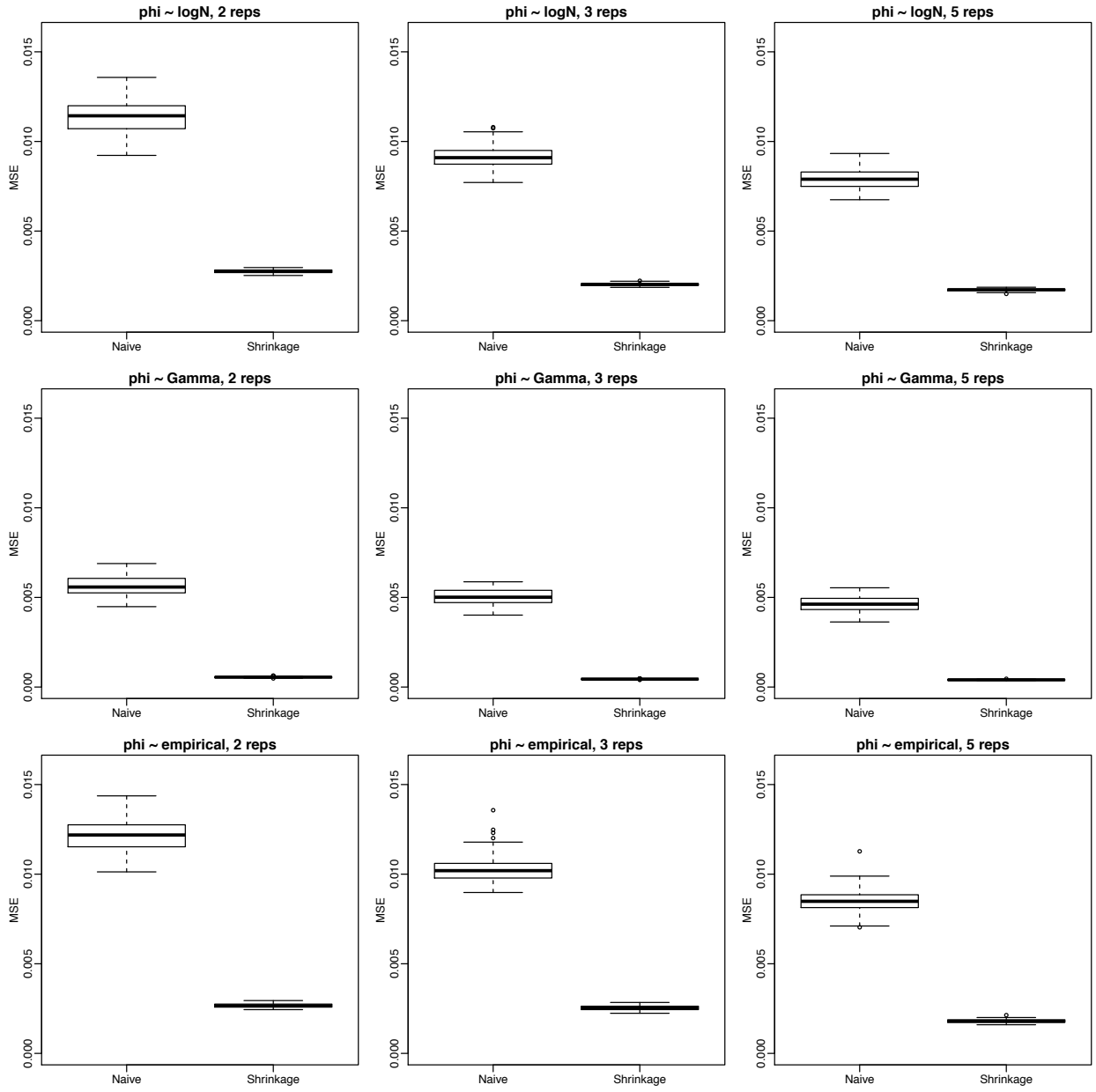
2. Comparison of dispersion estimation procedures

We compare dispersion estimates from a naïve method of moment estimator and the proposed shrinkage method based on simulations. The first two columns of Supplementary Figure 2 show the scatter plots of estimated versus true dispersions under different simulation scenarios. The third column compares the estimation biases from the two methods. The results show that the proposed shrinkage method achieves better accuracies and lower biases in dispersion estimation.



Supplementary Figure 2. Scatterplot of true dispersion vs. estimated dispersion from a naïve method of moment (MOM) estimator (left column) and the proposed shrinkage method (middle column), based on simulation studies. The right column shows boxplots of the estimation bias from one simulation. All plots are for 20,000 CpG sites and are based on 2 replicates per group (1st row), 3 replicates per group (2nd row) and 5 replicates per group (3rd row).

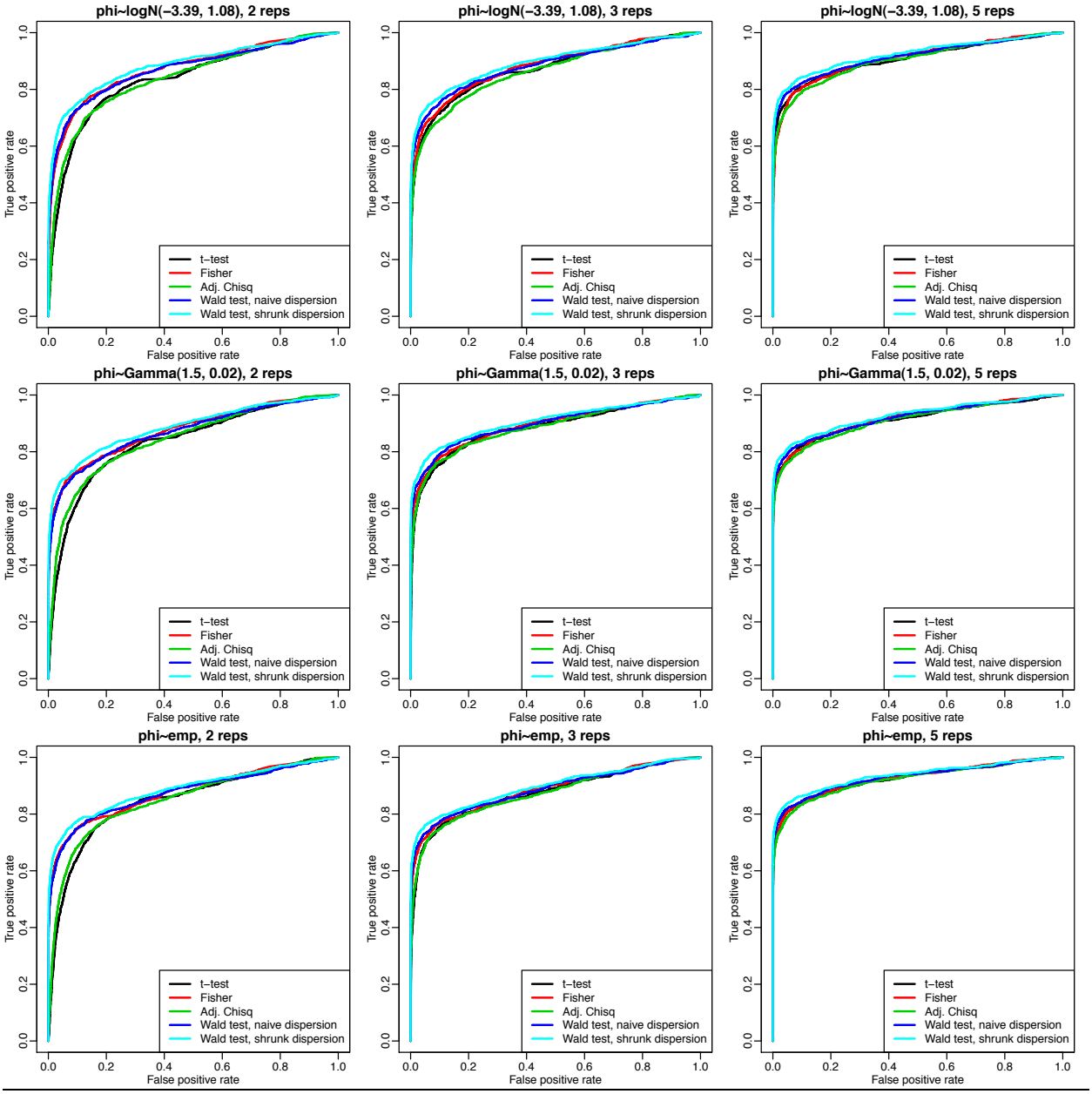
We further compare the MSE of the overall dispersion estimation biases from two methods based on 20 simulations. Supplementary Figure 3 shows that the proposed shrinkage method has a much lower MSE. Again this strengthens the point that the proposed method provides improved accuracy in dispersion estimation.



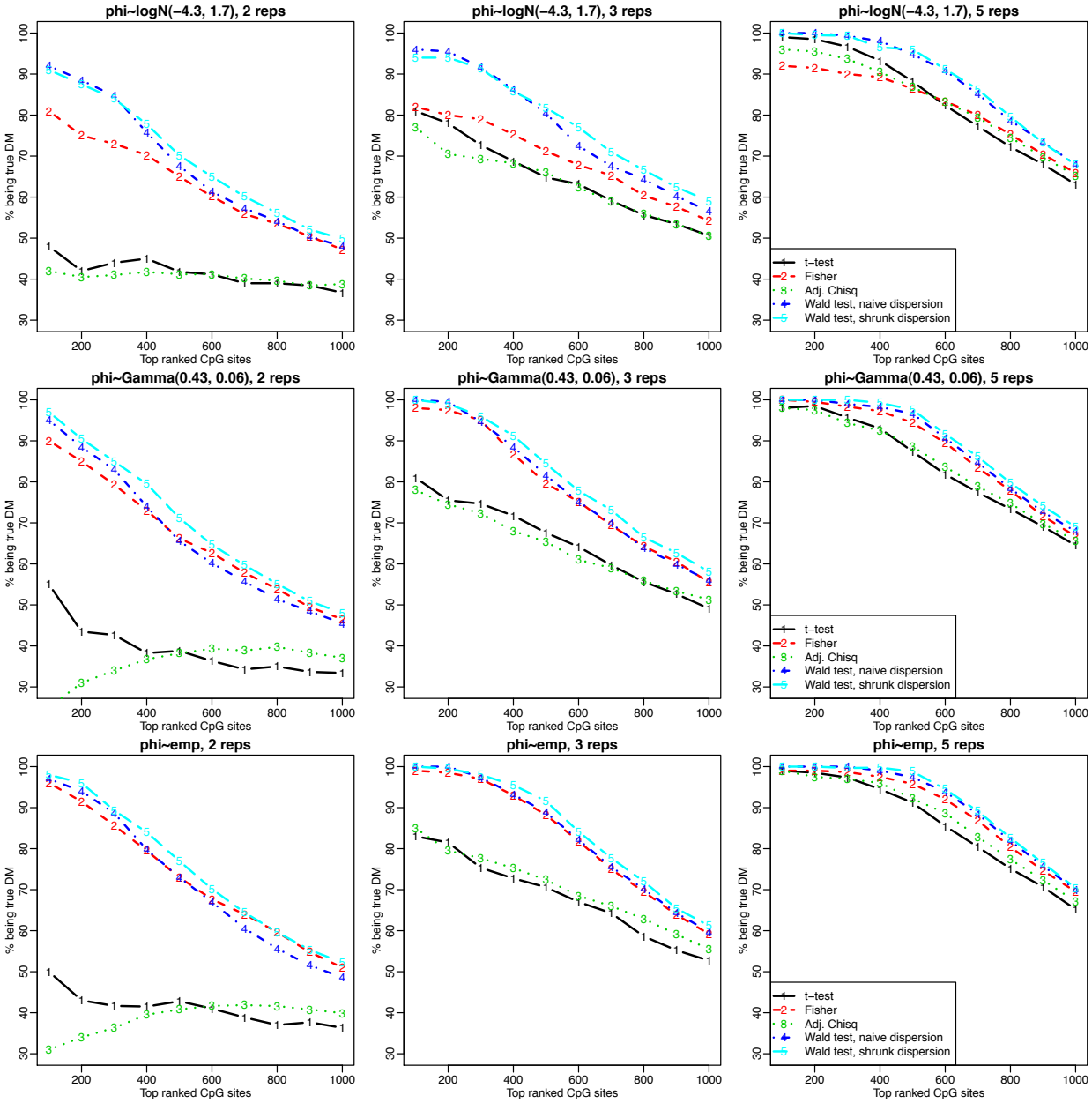
Supplementary Figure 3. Boxplots comparing the distribution of MSE for dispersion $\emptyset_i$ estimators from the proposed shrinkage method and naïve method of moment estimators (MOM) based on 100 simulations. Dispersion $\emptyset_i$ is randomly generated from the log-normal distribution (1st row), Gamma distribution (2nd row) and empirical distribution (3rd row). Each group contains 2 replicates (left), 3 replicates (middle) or 5 replicates (right). Each simulation is based on 20,000 CpG sites.

3. Comparison of DML detection accuracies—part B

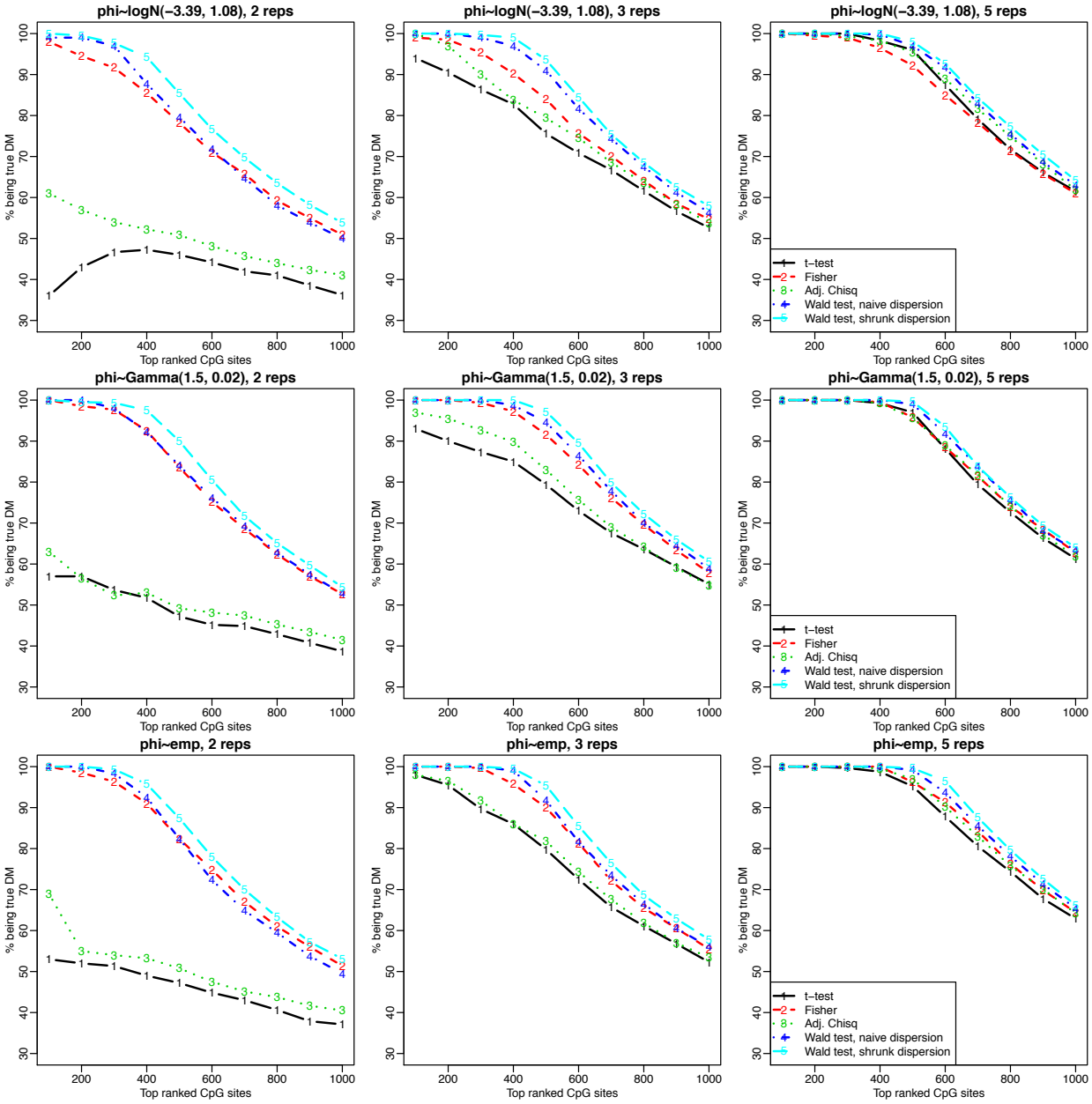
We provide results for additional simulations comparing the DML detection accuracies from different methods. Supplementary Figure 4 provides simulation ROC curves results based on data from the mouse genome (Smith *et al.*, 2012, Gene Expression Omnibus accession GSE34864), where the number of replicates and dispersion distributions used in the simulations vary. Supplementary Figure 5 provides another set of simulation results based on public data from the Arabidopsis genome (Gene Expression Omnibus accession GSE38991). Supplementary Figure 6 shows results from simulations where methylation levels across replicates follow a truncated normal distribution (instead of a beta distribution) at each CpG site. In all simulations, the Wald test with shrunk dispersion estimation achieves the best results, demonstrating that the proposed method is robust to different genomes and data distributions as well as model misspecification.



Supplementary Figure 4. The Receiver Operating Characteristic (ROC) curves of detecting true DMLs for different methods. True positive rate is plotted against false positive rate. The dispersion parameter $\emptyset$ comes from the log-normal distribution (1st row), Gamma distribution (2nd row), and empirical distribution from real data (3rd row). In situations where replicate numbers are low (1st column), our proposed method shows the largest improvement since it accounts for the difficulty of estimating variance from 2 replicates per group. Our proposed Wald test with shrunk dispersion estimator performs the best across all nine simulations.



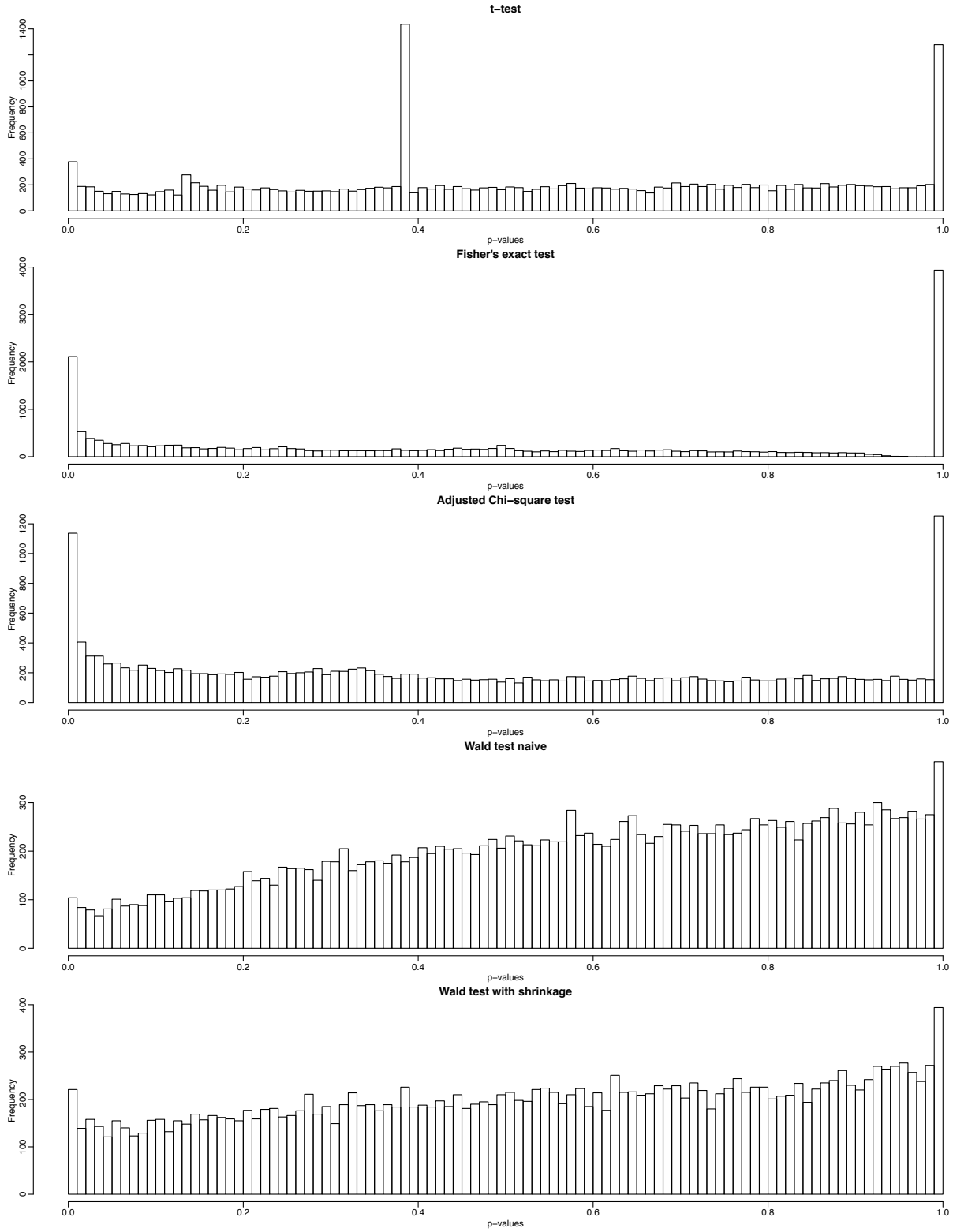
Supplementary Figure 5. The True Discovery Rate (TDR) plot of DML detection accuracy from simulations based on another genome. Parameters used in simulations are first estimated from a publicly available Arabidopsis genome dataset; simulation data is then generated under those estimated parameters. After testing for DML, CpG sites are ranked by p-value from low to high. The proportion of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. The dispersion parameter Ø comes from the log-normal distribution (1st row), Gamma distribution (2nd row), and empirical distribution from the real data (3rd row). In situations where replicate numbers are low (1st column), our proposed method shows the largest improvement since it accounts for the difficulty of estimating variance from 2 replicates per group. Our proposed Wald test with shrunk dispersion estimator performs the best across all nine simulations, even with different genomic information.



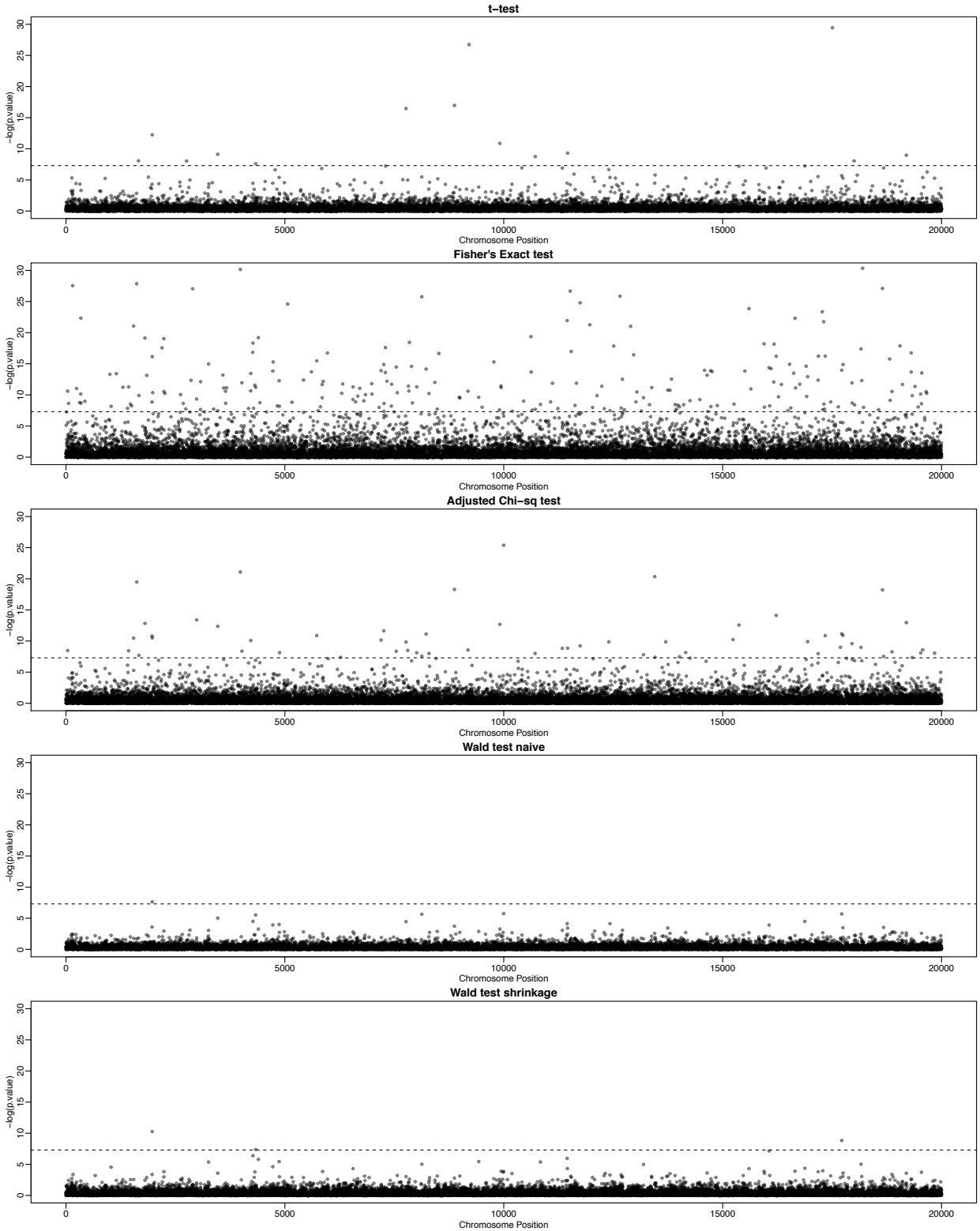
Supplementary Figure 6. The True Discovery Rate (TDR) plot of DML detection accuracy from using different distribution to generate methylation levels: methylation levels of biological replicates for each CpG site are generated from a truncated normal distribution. After testing for DML, CpG sites are ranked by p-value from low to high. The proportion of true discovery among top-ranked CpG sites is plotted against the number of top-ranked CpG sites. The dispersion parameter Ø comes from the log-normal distribution (1st row), Gamma distribution (2nd row), and empirical distribution from real data (3rd row). In situations where replicate numbers are low (1st column), our proposed method shows the largest improvement since it accounts for the difficulty of estimating variance from 2 replicates per group. Overall, our proposed Wald test with shrunk dispersion estimator continues to outperform the other tests, verifying the robustness of our proposed model to deviations from the assumed distribution.

4. Simulation under the null hypothesis

We use simulation to compare the results of our proposed method with existing methods. Simulations are based on a publicly available mouse genome dataset (Smith *et al.*, Gene Expression Omnibus accession GSE34864). For each simulation, we simulate 20,000 CpG sites for replicates from two groups. The simulated methylation levels are drawn from estimated values of μ from the mouse genome data. Dispersion parameters are simulated from log-normal distribution. We simulated under the null hypothesis where the underlying methylation levels for two groups are the same for each CpG site ($\mu_{i1} = \mu_{i2}$ for all i). We then tested for DML using each of the following approaches: t-test, Fisher's exact test, an adjusted Chi-square test, Wald test with naïve dispersion estimator and Wald test with shrinkage onto the simulated data. Supplementary Figure 7 shows the histogram of p-values from the 5 different methods under the null hypothesis. Supplementary Figure 8 is a Manhattan-style plot showing the distribution of p-values along the genome. These figures show that Fisher's exact test and adjusted Chi-square test yield a relatively high proportion of false positives in the simulated data.



Supplementary Figure 7. Histograms of p-values of all the tested CpG sites based on 5 different methods under the null hypothesis. Since the null hypothesis should be true for all CpG sites tested, one would expect p-values to follow a uniform distribution between 0 and 1. Our methods (Wald test framework) show a better goodness of fit to the uniform distribution (0, 1) [See **Table 1** for goodness-of-fit test statistics].

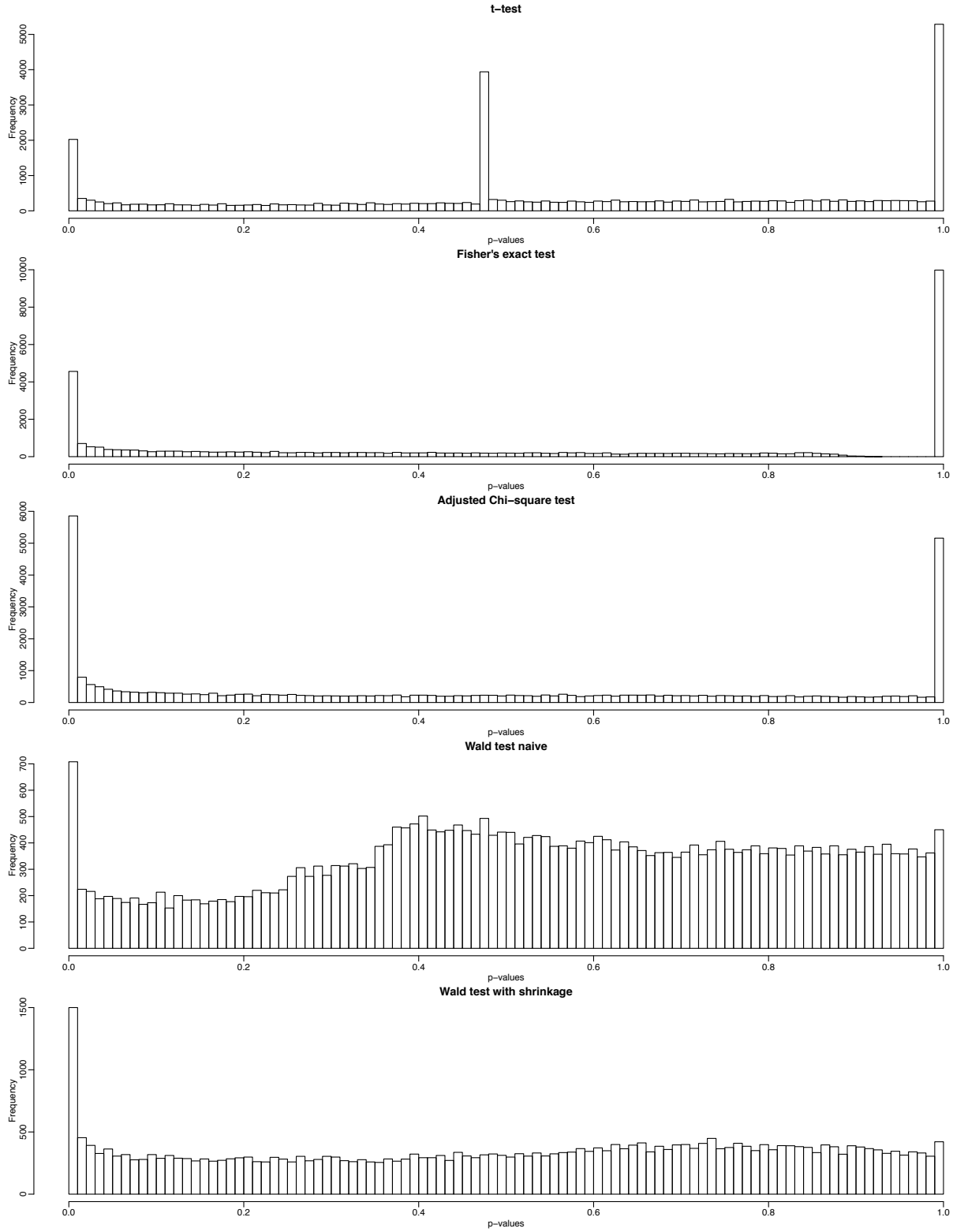


Supplementary Figure 8. Manhattan-style plot of logarithm of average of p-values from simulations under the null hypothesis. There are 4 replicates in each group in the 2-group comparison. Data of counts of methylated and total reads were generated under the null hypothesis ($\mu_{i1} = \mu_{i2}$). P-values were calculated for each CpG site; because we simulated under the null hypothesis, significant p-values (points above the dotted line) can be considered false

positives. Fisher's exact test and adjusted Chi-square test yield significantly more false positives when compared with the Wald test. Simulations were based on 20,000 CpG sites.

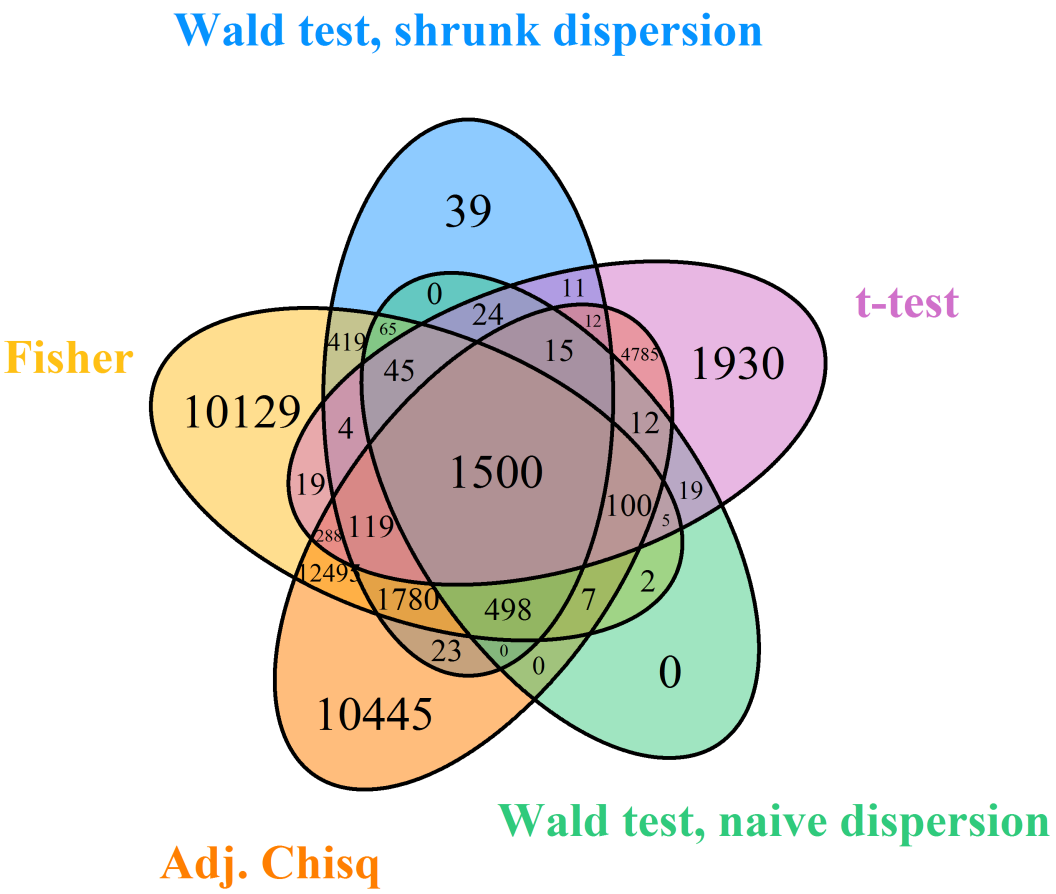
5. Distribution of p-values in real data

We analyzed methylation data from 2 replicates each of oocyte and zygote samples (Smith *et al.*, 2012). For all CpG sites on chromosome 1, we tested for DML between oocytes and zygotes using five methods: t-test, Fisher's exact test, an adjusted Chi-square test, Wald test with naïve dispersion estimator and Wald test with shrinkage dispersion estimator. P-values were recorded for all tested CpG sites. Supplementary Figure 9 shows the histogram of p-values for the five methods. Under the assumption that most of the CpG sites are not differentially methylated, only our proposed Wald test framework yields an appropriate mixture of uniformly distributed p-values (not significant) and p-values near 0 (significant). Supplementary Figure 10 shows the Venn diagram of whole genome detected DMLs from 5 different methods in real data analysis, where genomic Bonferroni-corrected p-value was applied as the cutoff to call DML. Large numbers of method-specific detected DML reflect a likely anti-conservative bias for those methods.



Supplementary Figure 9. Histograms of p-values of all the tested CpG sites based on 5 different methods. Since the null hypothesis should be true for the majority of CpG sites tested, one would expect most of the test p-values to follow a uniform distribution between 0 and 1, with a minority of test p-values near 0. Our methods (Wald test framework) show a good combination of the null hypothesis for a majority of CpG sites and alternative hypothesis

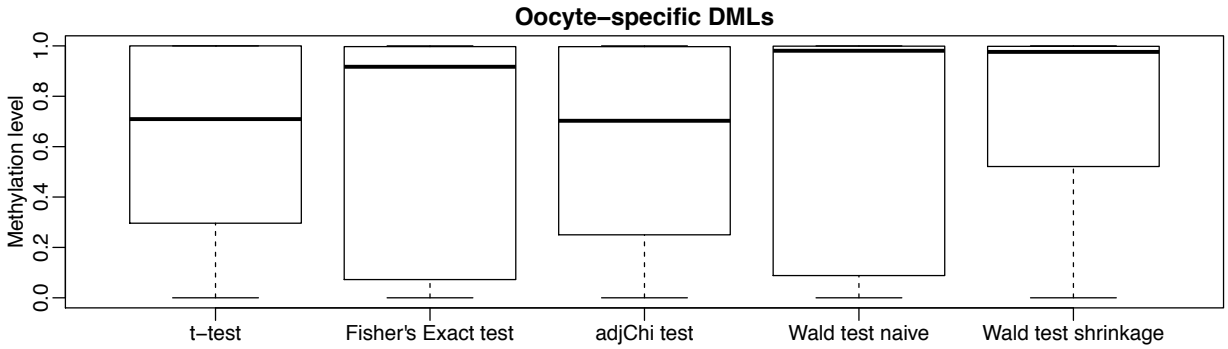
for a minority of CpG sites, while other methods show an excess of p-values near 0 and 1 (and near 0.5 in the case of the t-test).



Supplementary Figure 10. Venn diagram of detected DMLs from five different methods in oocyte vs. zygote samples analysis. Genomic Bonferroni-corrected p-value is applied as the cutoff to call DML. Large numbers of method-specific detected DML reflect a likely anti-conservative bias for those methods.

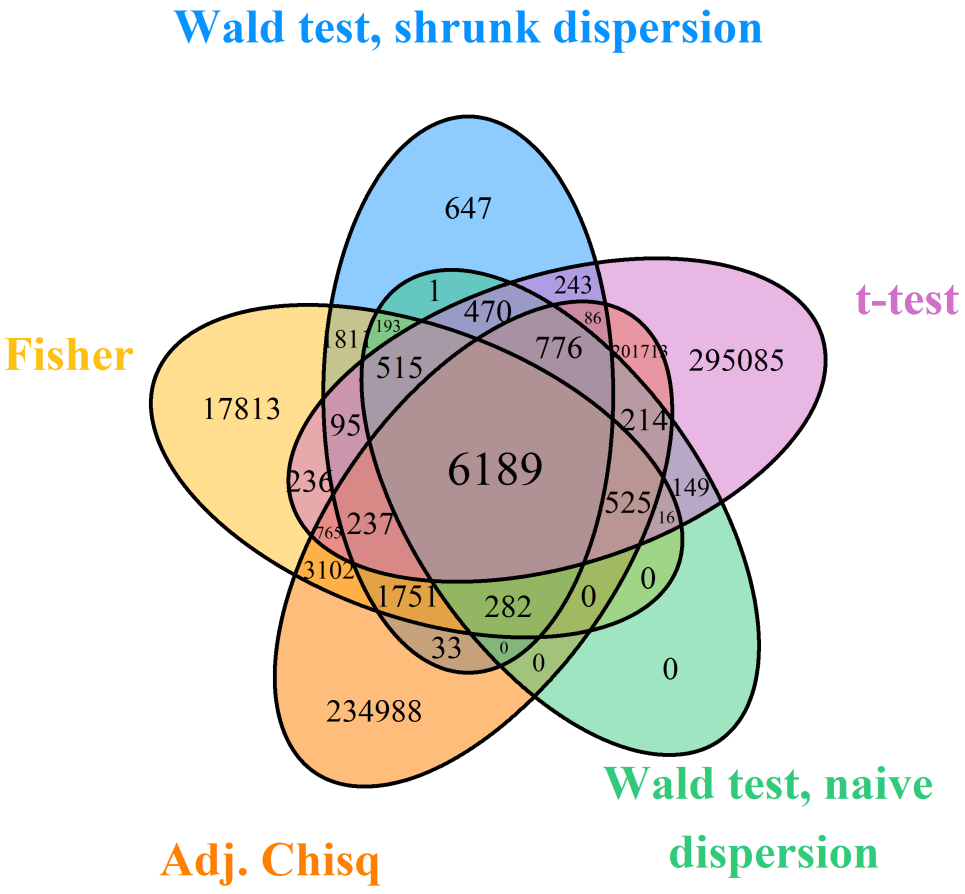
6. Concordance comparison with previous findings

We analyzed the same oocyte and zygote data via five different methods and compared our findings to those previously reported. We tested for oocyte/zygote DML in this dataset using t-test, Fisher's exact test, an adjusted Chi-square test, Wald test with naïve dispersion estimator and Wald test with shrinkage dispersion estimator. CpG sites identified as DML were sorted by p-value, from smallest to largest. The Smith *et al*. (2012) study reported that sites differentially methylated between oocytes and zygotes had methylation levels close to 1 in oocytes. Supplementary Figure 11 shows the boxplot of methylation level of the top 2,500 oocyte/zygote DML identified by each method. DML identified via our proposed Wald tests have the highest methylation levels (close to 1), suggesting a similar pattern to that reported by Smith *et al.*.



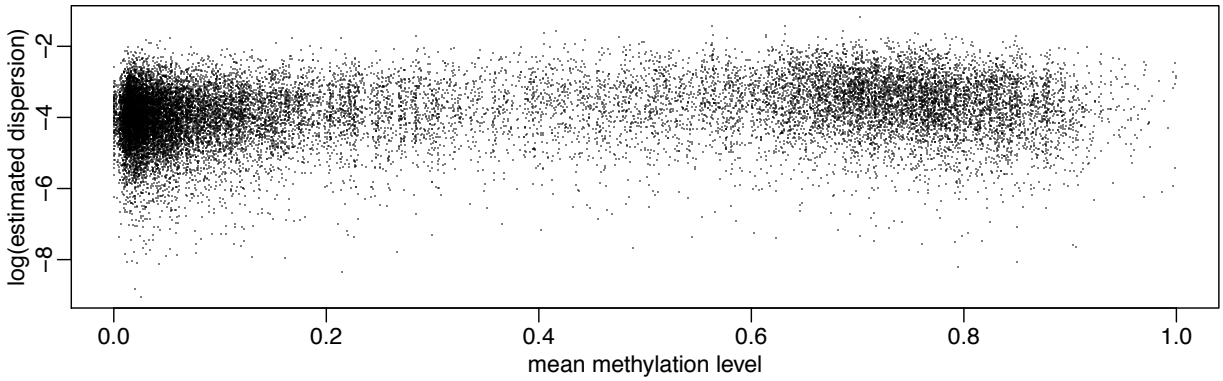
Supplementary Figure 11. Boxplot of methylation levels of oocyte-specific top 2,500 identified CpG sites from 5 different methods. Smith *et al.* (2012) reported that oocyte-specific CpG sites have methylation level close to 1. Here our proposed method has the highest concordance with those findings.

7. Venn diagram for analyzing the Alzheimer's disease vs. normal data



Supplementary Figure 12. Venn diagram of detected DMLs from five different methods in Alzheimer's disease vs. normal data analysis. Genomic Bonferroni-corrected p-value is applied as the cutoff to call DML. Large numbers of method-specific detected DML reflect a likely anti-conservative bias for those methods.

8. Dependence of dispersion and mean methylation levels



Supplementary Figure 13. Scatterplot of estimated dispersion in logarithm scale against estimated methylation level. Our method assumes the independence of dispersion of mean. This plot shows that there is no apparent trend in the relationship between the dispersion and mean, which would verify our assumption.

Supplementary Table

Supplementary Table 1. Hypothetical genes list associated with Alzheimer's disease. Gene names are obtained by mapping the identified DMRs with transcriptional start site (TSS) regions of gene on reference human genome.

REFSEQ GENE NAME	DESCRIPTION
PAX8, PAX8-AS1	Members of this gene family typically encode proteins involved in thyroid follicular cell development and expression of thyroid-specific genes (1). Thyroid hormone is involved in adult cognitive functions (2).
RGPD1, RGPD2 PLGLB1, PLGLB2	Associated with primate-specific duplications of genes (3) and alcohol and nicotine co-dependence (4).
LOC654433	An RNA gene, affiliated with the lncRNA class. No established functional reference.
FAM90A1	Interact with Amyloid Precursor Protein (APP) gene. APP is able to form the protein basis of the amyloid plaques found in the brains of patients with Alzheimer disease. Mutations in APP gene have been associated in Alzheimer disease (5).
LINC00654	Long intergenic non-protein coding RNA 654. No established functional reference.

Supplementary Material Reference

1. Mansouri, A., Chowdhury, K. and Gruss, P. (1998) Follicular cells of the thyroid gland require Pax8 gene function. *Nature genetics*, **19**, 87-90.
2. Beydoun, M.A., Beydoun, H.A., Kitner-Triolo, M.H., Kaufman, J.S., Evans, M.K. and Zonderman, A.B. (2013) Thyroid hormones are associated with cognitive function: moderation by sex, race, and depressive symptoms. *The Journal of clinical endocrinology and metabolism*, **98**, 3470-3481.
3. Ciccarelli, F.D., von Mering, C., Suyama, M., Harrington, E.D., Izaurralde, E. and Bork, P. (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res*, **15**, 343-351.
4. Zuo, L., Zhang, F., Zhang, H., Zhang, X.Y., Wang, F., Li, C.S., Lu, L., Hong, J., Lu, L., Krystal, J. *et al.* (2012) Genome-wide search for replicable risk gene regions in alcohol and nicotine co-dependence. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, **159B**, 437-444.
5. Matsui, T., Ingelsson, M., Fukumoto, H., Ramasamy, K., Kowa, H., Frosch, M.P., Irizarry, M.C. and Hyman, B.T. (2007) Expression of APP pathway mRNAs and proteins in Alzheimer's disease. *Brain research*, **1161**, 116-123.